Sensitivity of the population size estimates for Census undercoverage

Gerritse, S.C.

2016

Discussion Paper

Under coverage of the population register in the Netherlands, 2010

2016 | 02

Susanna C. Gerritse
Bart F.M. Bakker
Peter-Paul de Wolf
Peter G.M. van der Heijden (Utrecht University/University of Southampton)

Content

Summary

In this manuscript we consider an important topic in official statistics, namely estimating the number of usual residents. For the Netherlands we investigate the under coverage of the Population Register. First, the Population Register is linked to an Employment Register and a Crime Suspects Register. Then, we use three list capture-recapture methodology to estimate the number of usual residents. Two problems arise: 1) all three registers have no variable on usual residence; 2) capture-recapture methodology relies on a couple of assumptions, for which it cannot be verified whether they are met. The paper shows how the problems are solved.

1. Introduction

In this manuscript we are interested in estimating the under coverage of the Dutch Population Register (PR). The PR is an important register for Census purposes, and the undercoverage of this register gives an indication of the quality of the register. One common method is to link the PR to other registers and estimate the portion of the population missed by the registers using capture-recapture estimation, also known as multiple systems estimation (Fienberg, 1972; Bishop, Fienberg and Holland, 1975; Cormack, 1989; International Working Group I, 1995). Capture-recapture methodology is a general method to estimate the size of populations. An early example is estimating population specifics such as birth and death rates in an area near Calcutta, India (Sekar and Deming, 1949). The methodology has also been used regularly for Census purposes, for instance in the United Kingdom (Brown, Diamond, Chambers, Buckner, and Teague, 1999; Brown, Abbott and Diamond, 2006; ONS, 2012) and in the US (Wolter, 1986; Bell, 1993; Nirel and Glickman, 2009).

The capture-recapture methodology is well documented. The challenge, however, lies in the practical application of the methodology. For example, difficulties can arise in identifying which data sources best describe the population. Moreover, sometimes missing values are present in data on crucial variables, or assumptions of the method may be violated. Such difficulties can often not be avoided and solutions have to be found in order for the population size estimation to lead to correct outcomes.

Assessing the undercoverage of the Dutch Population Register asks for a definition of the Dutch population. Defining the rules according to which a person is, or is not, part of the population of a country has a lot of consequences, such as allocation of parliamentary seats in the EU and the attribution of funds depending on population size. Thus the definition of the population of a country is important statistical information and the Census is the primary framework to define a population (Lanzieri, 2013). According to the (United Nations Statistics Divisions, 2008) we can define the population of a European country along the terms of usual residence:

"1.461. In general, "usual residence" is defined for census purposes as the place at which the person lives at the times of the census, and has been there for some time or intends to stay there for some time"

According to the European Union, Regulation (EU) No 1260/2013 of the European Parliament, usual residence is defined as:

"The place where a person normally spends the daily period of rest, regardless of temporary absences for purposes of recreation, holidays, visits to friends and relatives, business, medical treatment or religious pilgrimage"

An individual is considered a usual resident when they have lived in the Netherlands for a continuous period of 12 months before a Census reference time, or if they

arrived in the 12 months before a Census reference time and intend to stay for at least a year. When these circumstances cannot be established, "usual residence" means the place of registered residence (European Parliament, 2008). In accordance with the European Union regulations, in this manuscript we use the definition of residing for 12 months for usual residence. However, intention to stay is not registered and instead we define usual residence as residing more than 12 months continuously in the Netherlands.

This manuscript will document the steps needed for the application of capture-recapture methodology on the case of the undercoverage of the Dutch PR using usual residents. In doing so, we document the problems that arose in achieving this research goal, and how they were handled. This research poses as an example of how to deal with possible problems in the practical application of capture-recapture methodology in Official Statistics.

The Dutch Population Register (PR) used for the 2011 Census round was still the Gemeentelijke BasisAdministratie (GBA [1]) Under Dutch regulations, every individual residing in the Netherlands for longer than four months, or is planning to do so, should register in the PR. As such, the PR contains demographic information on the 'de jure' population and differs from the 'de facto' population, which is the actual number of individuals residing in the Netherlands regardless of registration. The coverage of the PR alone is not sufficient to provide a valid estimate of the 'de facto' population, in this manuscript defined by number of usual residents.

This incompleteness of the PR has more than one reason. First, within the European Union there is free movement and employment for individuals with an EU nationality. When an individual with an EU nationality resides in the Netherlands for longer than four months without having registered in the PR, they are not illegal residents, despite that they can be fined by Dutch law for not registering. Individuals that have not registered themselves may have forgotten to do so, or simply do not want to. These individuals are considered usual residents by the definition of the European Union but belong to the undercoverage of the PR. Second, the PR is also incomplete due to immigrants, coming from outside the European Union without a working or residence permit. These individuals then are illegally residing undocumented immigrants. These illegally residing undocumented immigrants are also considered usual residents, but are part of the undercoverage of the PR.

The registered population will also contain an overcoverage, and this may occur when registered individuals no longer reside in the Netherlands because of, for example, administrative delay of registering emigration and death. In the Netherlands, however, this is not as big a problem as the under coverage. Bakker

[1] The GBA is currently replaced by the Basis Registratie Personen, (BRP). The BRP differs from the GBA because it also registers foreign individuals that have some sort of relationship with the Netherlands, and covers more individuals. This includes individuals with a Dutch nationality living abroad. Also, municipalities can register individuals that have not registered themselves, something which was not possible with the GBA. It is possible however is that the BRP, compared to the GBA, has a higher overcoverage of the Dutch population.

(2006) estimated an over coverage of 31 thousand individuals, which is only 0.2 percent of the PR registered population.

To estimate the number of individuals missed by the PR, we linked three registers: the PR, an Employment Register (ER) and a Crime Suspects Register (CSR). In the ER jobs are documented. In the CSR suspects of all known crimes are registered. Unfortunately we can easily deduce residence duration for the PR only. For the ER usual residence can in part be deduced based on job lengths, assuming individuals residing in the Netherlands during the period that they hold a job. Consecutive and overlapping jobs were considered as one residence duration. However, for individuals that are unemployed between jobs, a decision had to be made on a length of unemployment that will be allowed between two jobs to still be considered as one continuous residence duration. The CSR has no information to deduce residence duration and for those individuals not linked to the PR and/or the ER we use missing data methodology.

During the linkage process it was found that 37% of the individuals in the CSR that did not link to both the PR as well as the ER, had incomplete linkage key information. It may be possible that these individuals are also in one or both of the other two registers but could not be linked. A part of these individuals could be erroneous captures. When individuals that are suspected of a crime are not registered in the PR, the police cannot verify their information in the PR. Then, inaccurate information may occur, or even missing data. Therefore, when individuals have no crucial linkage key information, such as address and nationality, it is possible that these individuals do not belong to the population and, from our perspective, are erroneous captures. Gerritse, Bakker, Zult and van der Heijden (submitted) found that erroneous captures and linkage error can result in bias in the population size estimate resulting from capture-recapture analysis. In this paper, we propose different scenarios where, for the individuals in the CSR that cannot be linked and have missing data, the proportions of linkage error and erroneous captures are varied to assess the effect on the population size estimator.

Additionally, capture-recapture methodology relies heavily on a couple of assumptions that cannot be verified from the data (see for example: van der Heijden, Whittaker, Cruyff, Bakker and van der Vliet, 2012). In this manuscript every assumption of capture-recapture methodology will be discussed and we will use the information and resources available to meet these assumptions as best as possible.

The manuscript is structured as follows. In section 2 the linkage of the data sources used will be discussed. In section 3 we will discuss residence duration. First, in section 3.1, the sensitivity analysis for the residence duration of the ER will be discussed, followed by, in section 3.2, the missing data methodology used to impute the residence duration of the CSR. In section 4 we shortly discuss capture-recapture methodology and its assumptions of capture-recapture analysis and the scenarios we use, and it discusses how the analysis is conducted. Section 5 gives the results from the capture-recapture estimation. The results will be discussed and concluded in section 6.

2. Linkage of data sources

We use three registers in this manuscript, the PR, the ER and the CSR. For the capture-recapture analysis only individuals that did not have a Dutch nationality were considered. Individuals with a Dutch nationality were considered Dutch residents and were excluded from the analysis. This also included individuals who had two nationalities, of which one of them was Dutch. We use ultimo September 2010 as the reference time point, or reference date.

For the purpose of our analyses the ER has been transformed into a register on individuals, where jobs were attributed to the individuals holding those jobs. The ER adds new cases of usual residents to the PR because it also registers individuals working in the Netherlands that have not registered themselves in the PR. Individuals with an EU nationality are free to work in other EU countries, but for salary and tax purposes employers will register these individuals in the ER.

The CSR documents individuals that are suspected of a known crime. In principle, this register can hold information on everyone in the Netherlands, including undocumented immigrants or other non-PR registered individuals, because every individual residing in the Netherlands has a chance to become a suspect of a crime. Thus, it may provide cases that were not found in the PR but do belong to the population.

There is little information in the CSR and the ER on individuals with ages under 12, and over 65, because individuals under 12 cannot be registered in the CSR and the ER only registers between 15 and 65: As such it was decided that the population specified in this paper will consist only of the population aged between 15 to 65.

To link the three registers we used two types of linkage, deterministic and probabilistic, both of which are considered a type of exact matching (Herzog, Scheuren and Winkler, 2007). For a pair to be a link under deterministic linkage, the two records have to agree exactly on each element in a set of identifiers. One example is when two records match on name, address, date of birth and city of birth, or on a Personal Identification Number. Fellegi and Sunther (1969) formalized probabilistic linkage where all records in one register are paired to all records of a second register. Based on their agreement on a set of identifiers, a weight is given to each pair. A predefined cut-off determines whether pairs are links, non-links or possible links (Herzog, Scheuren and Winkler, 2007). To reduce the number of possible pairs for computational efficiency we can block the data on one, or more variables.

At Statistics Netherlands all registers are linked deterministically to the PR via a PIN and a linkage key. The PR is the backbone of Statistics Netherlands, such that all registers and surveys are linked to the PR. Linking other registers, of which also the ER and the CSR, to the PR via a Dutch PIN, enables about 96 to 98 percent of the

cases to be linked to the PR. When a case has no PIN, the registers are linked on postal code, house number, date of birth and sex. Then 93 to 95 percent of individuals can be linked to the PR (Arts, Bakker and van Lith, 2000). Thus it seems that when cases have complete linkage keys, linkage to the PR can be done in more than 90 percent of the cases. However, there are cases with incomplete linkage key information, such that they cannot be linked. These cases are often without a Dutch nationality and are part of the subpopulation of interest for this study. To improve upon the deterministic linkage we also used probabilistic linkage.

For the individuals in the PR and ER we used a combination of date of birth, sex, postal code, house number and suffix as linkage key to probabilistically link the remaining ER units to the whole of the PR. This resulted in an overall improvement of 0.1% of the remaining ER-units that can be linked. The same linkage key is used to probabilistically link the remaining units in the CSR to the whole of the PR, which led to an overall improvement of 3.3% of extra individuals linked. For both the linkage of the ER to the PR and the CSR to the PR we blocked first on postal code or date of birth to reduce the number of possible pairs. To probabilistically link the whole of the ER to the whole of the CSR we used a combination of birth date, sex, city of residence, country of residence, street name and house number as linkage key. For the probabilistic linkage of the ER to the CSR, the data was blocked on either city of residence of birth date had to be equal to at least day or month of birth and led to an overall improvement of 9.9% of extra individuals linked.

Figure 1 shows the linkage of the PR to the ER, the PR to the CSR and the ER to the CSR, and more specifically the percentage of each register linked either deterministically (in yellow) or probabilistically (orange). It was found that 37.7% of the individuals in the CSR that could not link to the PR or the ER had incomplete linkage key information. Thus these individuals were unable to be linked.

We use four covariates for capture-recapture analysis: nationality group, age, sex and usual residence. Nationality group has 7 categories: (1) EU15 (excl. Netherlands) (2) Polish (3) Other EU (4) Other western (5) Turkish, Moroccan, Surinam (6) Iraqi, Iranian, Afghan, asylum seeker countries Africa (7) Other Balkan, former Soviet Union, other Asian, Latin American and other nationalities. The countries are clustered according to likely migration motives, migration legislation, regulations of the PR and size. For age, we use four levels: (1) 15-24 (2) 25-34 (3) 35-49 and (4) 50-65.

2.1.1 Observed values for the three registers

PR	ER	CSR		Total
		Yes	No	
Yes	Yes	2,108	259,811	261,919
	No	4,859	350,554	355,413
No	Yes	354	112,530	112,884
	No	5,086	0	5,087
	Total	12,407	722,895	735,302

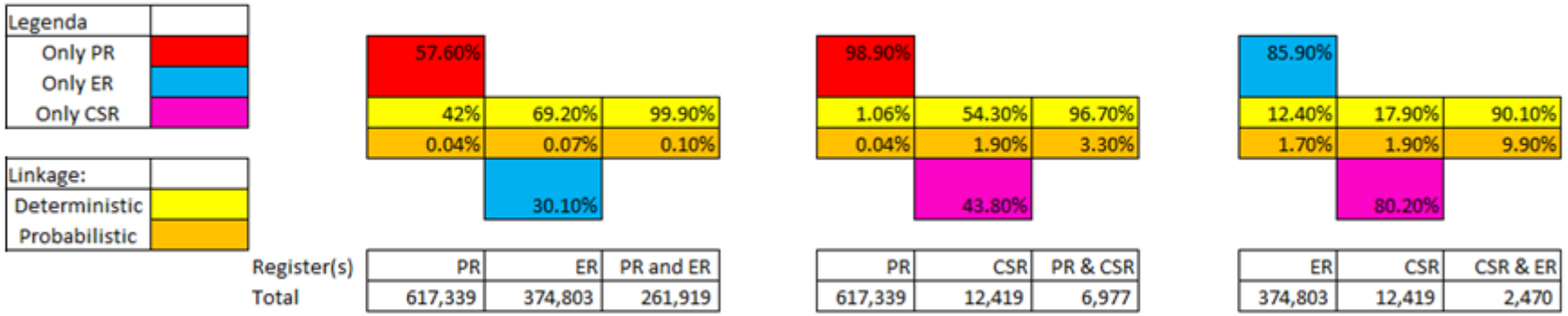

Figure 1. Linkage of the PR to the ER, the PR to the CSR and the ER to the CSR. Of the first figure, the first column shows all the individuals in the PR, where the second column shows all the individuals in the ER. The overlapping parts in yellow represent the percentage of individuals that have been linked deterministically, and the parts in orange represent the percentage of individuals that have been linked probabilistically. The last column shows the linked part between the PR and the ER, where it can be seen that probabilistic linkage led to an improvement of 0.1%. The other two figures show the same kind of information for respectively the PR linked to the CSR, and the ER linked to the CSR.

Table 2.1.1 shows the counts for the individuals in the three linked registers ignoring the distribution over the four covariates, as was also used in Gerritse, Bakker and van der Heijden (2015). Table 2.1.1 shows that the individuals are not evenly distributed over the three registers, which may lead to small cell counts in the capture-recapture analysis when the four covariates are taken into account. This may complicate the population size estimation. For that purpose, we use different scenarios to account for the biggest variability in this manuscript, and assess the impact on the population size estimation.

3. Methods for deriving residence durations

Neither register has a direct measure of residence duration. For the PR a measure of residence duration can be deduced. The PR has a registration date, the date at which either a person was born into the Netherlands or when immigrants registered themselves in the PR as a Dutch resident. Then the reference date of ultimo September 2010 minus the registration date can be used as a residence duration. Additionally, when a PR registration was consecutive or overlapping with a job, such as when a job was registered in the ER before the PR registration, the combined length of PR and ER registration was also considered a residence duration. The following sections will document the steps taken for the usual residence measure of the ER and the CSR.

3.1 Residence duration ER

In the ER, amongst others, the start date and the end date of the job are registered. Assuming that individuals that have a job in the Netherlands, also reside in the Netherlands during the time they hold that job, we can use the information on successive jobs previous to the reference date as a residence duration. For individuals that held more than one job that were consecutive or overlapping we perceive these jobs as one period of work and therefore residence. A total of 77% of the 374,803 individuals in the ER had one period of residence.

When there was a period of unemployment between two jobs, the question is whether this affects an individual's usual resident status. For the individuals in the ER that did not link to the PR, and thus did not have a residence duration from the PR, a decision had to be made on the length of unemployment between jobs that would be allowed, wherein the individual still perceived as residing in the Netherlands. We investigated seven scenarios. A total of 1, 8, 15, 22, 31, 62 and 93 days were allowed between two jobs for an individual. In the first column of Table 3.1.1 are the observed values of all individuals ER by nationality groups. The other columns show for each of the scenarios the observed number of individuals in the ER but not in the PR that reside in the Netherlands for longer than a year. As can be seen, for an individual with a EU15 nationality 7,781 individuals are considered residing in the Netherlands for longer than a year based on their job length when we accept an unemployment of 8 days between two jobs. If we increase this to 15 days the number of individuals residing in the Netherlands increases to 7,915, up to 10,861 when we consider 93 days. The different scenarios give different numbers of usual residents.

3.1.1 Scenarios for deducing residence duration from job-lengths in the ER, per nationality group. The first column shows the observed counts of all the individuals in the ER that are not in the PR. The other columns show the count of individuals that would be considered usual residents under the specified scenarios.

Nationality	Total	1 day	8 days	15 days	22 days	31 days	62 days	93 days
EU15	18,727	7,548	7,781	7,915	8,093	8,862	9,982	10,861
Polish	80,738	14,711	15,677	16,301	16,904	20,315	25,520	29,399
Other EU	10,765	2,201	2,267	2,331	2,369	2,617	3,095	3,42
Other West	509	216	218	220	221	222	231	242
Turkey, etc.	647	341	349	354	363	380	425	457
Iraq, etc.	274	142	145	149	155	166	179	194
Balkan, etc. Other.	1,378	530	539	543	553	567	615	671
Total	113,038	25,631	27,034	27,871	28,717	33,193	40,126	45,325

It is difficult to choose which scenario will be more realistic. As such a choice has to be made on the few indications that can be found in Table 3.1.1. We have chosen for the scenario of 31 days. A choice could be made for a smaller scenario, but it seemed more realistic to allow for one month, given that jobs often start at the beginning of a month and end at the end of the month. Thus allowing at least a month is plausible. Also, it can be seen from Table 3.1.1 that the biggest, absolute, increase in usual residents is between 31 and 62 days, which indicated a turning point between these scenarios where more individuals may decide to leave the country for a longer period. The biggest group of non-Dutch individuals is of European nationality, and the probability that they return to their home country after 31 days increases.

Additionally, it is slightly more conceivable that when individuals are unemployed for up to 31 days that they are still in the Netherlands, compared to when more months are allowed. A total of 113,038[2] individuals in the ER are not in the PR, of which we consider 33,135 individuals are usual residents. The remaining 79,903 individuals in the ER that did not link to the PR are non-usual residents. Note that for the individuals in the ER that are also in the PR, we have two residence durations. One for the ER registration via job lengths and one from the PR. From both residence durations we chose the residence duration that is longest.

3.2 Residence duration CSR

For the CSR it is not possible to define a residence duration based on the variables in the register itself. Moreover, the CSR is an event based register in which crimes are recorded on one particular date. Most of the suspects are first offenders, which

[2] After probabilistic linkage of the ER and PR to the CSR it was found that 154 ER registered individuals, and 7 PR registered individuals, were erroneously seen as CSR registered individuals only.

means that there is only one date recorded for most individuals and a residence duration cannot be deduced from this one date. For the records that link to the PR and/or the ER, we use the residence duration from these sources. When usual residence was available from both the PR and ER, the longest residence duration of either the PR or the ER was chosen. However, for the remaining records in the CSR, residence duration is missing, and we have to impute it.

In earlier research it was found that the missing residence duration values for the CSR will probably resemble the usual residence values from the ER more than the usual residence values from the PR (Gerritse, Bakker, van der Heijden, 2015) The CSR records will resemble the ER records rather than the PR records, because the individuals that register themselves in the PR have the intention to stay for a longer period in the PR. This cannot be assumed for the individuals in either the ER or the CSR that do not link to the PR, given there is a reason why these individuals have not registered themselves. As such we need an imputation method that imputes missing data with this very specific subset of data in the ER that it resembles most. It was found that for our specific missing data problem, multiple imputation by Predictive Mean Matching (PMM) will most likely handle the missing data best, given that PMM allows the user to ignore individuals in the PR (Gerritse, Bakker, van der Heijden, 2015)

Predictive Mean Matching is an example of a hot deck, nearest neighbour multiple imputation method. For each missing entry, PMM samples a small set of candidate donors from the observed complete cases (van Buuren, 2012). The predicted values are estimated based on the information in the predictor variables and their "closeness" is estimated by their absolute difference. From this small set of cases, one is randomly chosen to replace the missing value. The assumption here is that the missing value follows the same distribution as the observed values chosen in the small set (van Buuren, 2012).

This means that we use all 113,038 records from the ER and CSR that did not link to the PR to impute the missing usual residence value for the individuals in the CSR only. These donors will be used to impute the missing residency variable in the CSR units that do not link to either of the other sources. PMM has been repeated ten times, to account for multiple imputation variability. The PMM multiple imputation has been done via the package mice in R (van Buuren).

4. Method for estimating the number of unregistered individuals

Capture-recapture methodology is an often used methodology for population size estimation. The simplest form of capture-recapture methodology consists of linking two sources of data, such as samples, lists or registers. Assume we have two registers. Record linkage finds cases in both registers that are from the same individual by using identifying variables, such as a PIN or a combination of linkage key variables like name, address, etc. The result is a count of individuals in either register alone or in the overlap of both registers. Additionally there is a zero count of individuals that do belong to the population but were not observed in either register.

Assume we have two registers, register 1 and register 2. Let *i* = (0, 1) respectively denote not included in register 1 and included in register 1. Also, let *j* = (0, 1) respectively denote not included in register 2 and included in register 2. Let m_{ij}denote the expected values for registers 1 and 2, and n_{ij}denote the observed counts. Then odds ratios can be used assuming independence between register 1 and register 2, such that $m_{11}m_{00}/m_{01}m_{10} = 1$. However, m_{00} is a structural zero and is the value we are interested in. Assuming independence, this odds ratio can be rewritten to get maximum likelihood estimate

$$\hat{m}_{00} = \frac{\hat{m}_{01}\hat{m}_{10}}{\hat{m}_{11}} = \frac{n_{01}n_{10}}{n_{11}} \qquad (1)$$

This odds ratio is for two registers only, but can be easily extended from Equation (1) to the three register case. Let m_{ijk} be expected values for registers 1, 2 and 3. Let variable C denote inclusion in register 3, such that k = (0, 1) respectively denotes not included in register 3 and included in register 3. Then odds ratios are used assuming the three factor interaction to be absent, so that $m_{110}m_{000}/m_{101}m_{011} = m_{111}m_{001}/m_{100}m_{010}$. Expected value m_{000} is a structural zero and can be estimated by

$$\hat{m}_{000} = \frac{\hat{m}_{010}\hat{m}_{001}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{110}\hat{m}_{101}} = \frac{n_{010}n_{001}n_{100}n_{111}}{n_{011}n_{110}n_{101}} \qquad (2)$$

Equations (1) and (2) assume saturated loglinear models and can be readily implemented. More parsimonious models, compared to the saturated model, are possible and may fit the data better. Such models will be used in capture-recapture estimation later in this manuscript. For equation 2 using more parsimonious models leads to the adjustment that the last step, where fitted values are equated to observed counts, cannot be made.

The capture-recapture analysis has been conducted via Generalized Linear Modeling (GLM) in R. The function STEP in R enables the researcher to select the best fitting

loglinear model. By default STEP selects models based on the AIC. However since our sample size is quite large, the AIC will lead to models that are unnecessarily complicated. Therefore we used the BIC, since the BIC has a larger penalty for sample size. The details on how the variance for confidence interval testing was estimated can be found in the Appendix.

A summary has been made of previous research on the past number of unregistered individuals in the Netherlands in (Gerritse, Bakker and van der Heijden, 2015). We present here only the conclusion of this summation, as the arguments are presented in that manuscript. We expect that the under coverage of the PR will lie within a range of 175 to 225 thousand individuals. This range has been deduced from research by Hoogteijling (2002), Bakker (2006) and van der Heijden, Cruyff and van Gils (2011), and migration flows and asylum requests since these research articles. We acknowledge that this range is based on a couple of assumptions that may not be met, and we use this range with caution as an indication where the undercoverage may lie. Thus possible estimates of the under coverage of the PR that falls outside this range is not necessarily false. However, when our estimates lie far from the lower or upper boundary of this range, the estimate does become rather implausible.

4.1 Implied coverage

Capture-recapture relies heavily on a couple of assumptions (van der Heijden, Whittaker, Cruyff, Bakker and van der Vliet, 2012) Violation of these assumptions could lead to biased estimates, as was found in Brown, Abbott and Diamond (2006), Boden (2014), Gerritse, van der Heijden and Bakker (2015) and Gerritse, Bakker, Zult and van der Heijden (submitted). These researchers found that sometimes violated assumptions have little effect on the population size estimate, whereas sometimes it had a large impact on the population size estimate. The effect a violated assumption has on the population size estimator has been found to be a direct result of the implied coverage of the main register (Gerritse, Bakker, Zult and van der Heijden, submitted). Implied coverage describes the observed coverage of register 1, given register 2. As such, it describes the number of new cases added by register 2, compared to the already known cases in register 1.

From Equation (1) we can estimate conditional probabilities $\hat{p}_{(0|1)} = n_{01}/n_{+1}$ and, $\hat{p}_{(1|1)} = n_{11}/n_{+1}$.Thus $\hat{p}_{(0|1)}$ is the estimated probability of new cases from register 2, among all cases in register 2, and $\hat{p}_{(1|1)}$is the estimated probability of already known cases from register 1, among all the cases from register 2. Then Equation (1) changes to

$$\widehat{m}_{00} = \frac{\hat{p}(0|1)n_{10}}{\hat{p}(1|1)}. \qquad (3)$$

When $\hat{p}(0|1)$ is relatively small compared to $\hat{p}(1|1)$ the effect of the added new cases of register 2 on $\widehat{m}_{00}$will be small and the population size estimator is robust to violations of the assumptions. However, when $\hat{p}(0|1)$ is relatively large compared to $\hat{p}(1|1)$ the effect of the added new cases of register 2 on $\widehat{m}_{00}$ will be large as well

and the population size estimator is not robust to possible violations of the assumptions.

4.2 Implied coverage for three registers

Implied coverage can be extended to the three register case. Assume that register 1 covers most of the population, then register2 and then register 3. For our purposes it suffices here to discuss coverage of registers 1 and 2 implied by the third register. For this purpose Equation (2) is complicated, as n_{111}, the number of cases seen in all three registers, is in the numerator. We focus on the observed counts n_{101}, n_{011} and n_{001}, i.e. the number of individuals seen only in register 3, i.e. n_{001}, compared to the individuals seen in register 3 and only in register 1, i.e. n_{101}, and compared to the individuals only seen in register 3 and only in register 2, i.e. n_{011}. We focus first on n_{001} and n_{101}. Notice that these counts refer to individuals missed by register 2. Therefore we will speak of coverage conditional on being missed by register 2. Now, similar to the discussion of implied coverage in a two-way table above, if n_{001} is large in comparison to n_{101}, the conditional coverage of register 1 implied by register 3 is low, where conditional refers to being missed by register 2. Thus the estimator in equation (2) becomes unstable when the number n_{001} becomes unstable. Similarly, for n_{001} and n_{011}, if n_{001} is large in comparison to n_{011}, the conditional coverage of register 2 is low and the estimator in equation (2) becomes unstable when the number n_{001} becomes unstable (where conditional refers to being missed by register 1). Anyhow, for all practical purposes it will be clear that when n_{001} is large, the estimator defined in (2) will become unstable.

Table 2.1.1 reveals that the implied coverage of the PR, given the ER is relatively high. The implied coverage of the PR, given the CSR however is rather low. The implied coverage of the ER, given the CSR is also low. Summarizing, the conditional coverage implied by the CSR is low. This indicates that the population size estimator is not robust to possible violations of the assumptions.

4.3 Assumptions

Capture-recapture analysis relies heavily on a couple of assumptions of which cannot be verified from the data whether they are met. Above we stated that implied coverage seems low and thus the estimator may not be robust to possible violations of the assumptions. As such we discuss below the extra steps taken to make sure the assumptions are met as best as possible.

4.3.1 Closed population

The first assumption is that the population is closed. For the PR and the ER this assumption can be easily met. The PR and ER contain information on individuals over a period of time, such that one time point can be chosen to keep a closed population.

In our case the time point chosen is the reference date of ultimo September 2010. The CSR however is a register containing solely reports on suspects of all known crimes, and one specific day cannot be chosen since on that day only a limited number of reports could have been filed.

Thus a specific time period has to be chosen in which observations are taken from the CSR. This makes it difficult to assume a closed population. For the CSR a time period of half a year, the second half of 2010, has been chosen. We have chosen for half a year because then the number of individuals in the CSR is still relatively large, but also this time period is short enough so that a violation of the closed population assumption will be relatively minor. The second half of 2010 has been chosen because then the reference point used for the PR and ER lies in the middle of the six months considered.

4.3.2 Independence and homogeneous inclusion probabilities

Under independence one assumes that the probability to be included in the first register is independent on the probability to be included in the second register. This is a rather strict assumption. There are two ways to relax the independence assumption. One way is in using multiple sources. In using three registers the strict independence assumption is relaxed into the assumption that the three way interaction is zero. Additionally, in adding covariates heterogeneity due to these covariates can be removed (Compare, International Working Group I, 1995) In this manuscript we use three registers and four covariates, i.e. age, sex, nationality group and residence duration, to relax the strict independence assumption and account for possible heterogeneity in these covariates.

It has to be noted that the covariate usual residence is operationalised differently per register. Additionally, when cases were in the overlap of two registers, the longest usual residence value was chosen, and as such the operationalisation of usual residence in the overlap of registers is different to the operationalisation of the individual registers as well. Then dependence may well be introduced in usual residence considering the registers. In using usual residence as a covariate in the loglinear model in estimation, this dependence can be accounted for.

4.3.3 Perfect linkage

In this manuscript two methods for linkage have been used to increase the probability that we linked all cases that had to be linked. At Statistics Netherlands every register is linked deterministically to the PR, such that the PR and the ER, and the PR and the CSR, were already linked deterministically. However, due to errors in the variables used for linkage of the CSR and ER it was difficult to deterministically link the CSR to the ER. To improve the deterministic linkage, probabilistic linkage was used (see section 2 for more details).

4.3.4 No erroneous captures

The last assumption considered here is that the registers only contain information on the specified population, and thus do not contain any erroneous captures. When registering in the PR an individual has the intention to stay for a longer period, and thus they are assumed to be usual residents. Then for the PR we assume that there will be no erroneous captures. Individuals with an address in Belgium or Germany are removed from the analysis, as we assume that individuals from our neighbouring countries will probably still live in Belgium and Germany and only cross the border for work, school or possibly crime related activities. Additionally individuals are removed that were caught by the border police and thus did not enter the Netherlands at all.

4.3.5 Concluding remarks

It was found that 37 percent of the individuals in the CSR that did not link to the PR and the ER had missing or incomplete linkage key information. There is a chance that these individuals do belong to the population and had to be linked to the PR and the ER. There is also the possibility that these individuals have missing information for the reason that they do not belong to the population and thus are erroneous captures. Unfortunately, we cannot deduce from the data to which extent the assumption that there are no erroneous captures is violated. Additionally, we found that implied coverage is low due to a large number of individuals in the CSR that cannot be linked to the PR and the ER, and as a result the population size estimator will not be robust to possible violations of the assumptions. For that reason, we will investigate via scenarios which percentage of linkage errors and erroneous captures seem most realistic.

For the individuals in the CSR that did not link to the PR and the ER there is still the probability of duplicate cases. Due to the incomplete linkage key information we cannot establish the number of duplicates. We can however treat such cases as erroneous captures, and investigate the effect of removing these individuals from the analysis.

4.4 Scenarios

4.4.1 Main scenarios

Because implied coverage for the three registers is low, we use different scenarios to assess the effect of the presence of linkage error and erroneous captures in the 37% of individuals in the CSR that had incomplete linkage key information on the population size estimate. A baseline scenario will be set up where we assume no linkage error and no erroneous captures, and consider all 37% to belong to the population and to have only been registered in the CSR. For scenario 1 and 2 we will divide up the 37 percent of individuals in the CSR without linkage key information as either linkage error or erroneous captures.

The probability is higher that the individuals with incomplete linkage key information do not belong to the population and are more likely to be in the Netherlands as a tourist or for criminally related purposes. The police uses the PR to identify suspects.

When the police cannot find suspects in the PR, they will denote whatever information is available on the suspect. It is plausible that the police cannot denote information for individuals that do not belong to the population. When these individuals have no Dutch residence it will be a hard task to procure such information, resulting in missing information.

Methodologically this argumentation can be supported by the recent investigations of Zhang and Dunne (2015) and Zhang (2015) who researched capture-recapture methodology in the presence of erroneous captures. A trimmed dual system estimation method was set up. Erroneous captures are identified by linking the register to a post enumeration survey (PES). By assuming that the PES does not have erroneous captures, the size of the over coverage due to erroneous captures can be estimated. The next step is that records are to be excluded from capture-recapture analysis. The trimmed dual system estimator reduces the bias caused by erroneous captures by deleting them from the analysis. However, identifying erroneous captures is a hard task and the researchers advise against randomly deleting cases. Knowing that the 37 percent have no linkage key information, removing most of these individuals is a first step in deleting erroneous captures.

Therefore, we assume that for scenario 1 and 2 the majority of the 37 percent are erroneous captures. In scenario 1 we will consider a random selection of 75% of the individuals without linkage key variables as erroneous captures and remove them from the analysis, and the other 25 percent will be considered as linkage errors and these individuals from the CSR that did not link to the PR and ER will be linked to the PR and ER. Scenario 2 will investigate the effect of removing all individuals without the linkage key variables, the total of 37 percent, as though they are all erroneous captures.

4.4.2 Subscenarios

Additionally, there is a possibility that of the 63 percent of individuals in the CSR that did not link to the PR and ER, and do have complete linkage key information, there still are some linkage errors or erroneous captures. The CSR might contain possible administrative errors that have to be dealt with, even though these errors will be small. As such, scenarios 1 and 2 will have four possible outcomes, as shown in Table 5.1.1. Subscenario a from Table 5.1.1 (scenarios 1a and 2a) will have no linkage error or erroneous captures taken from the 63% individuals with complete linkage key variables. This will provide a baseline estimate for the scenario. In subscenario b (scenarios 1b and 2b from Table5.1.1an extra 5 percent of erroneous captures will be removed from the individuals with complete linkage key information and in subscenario c (scenario 1c and 2c from Table 5.1.1 an extra 5 percent of linkage errors will be taken from the individuals with complete linkage key information. In subscenario d (scenario 1d and 2d from Table 5.1.1 both an extra 5 percent of erroneous captures and linkage errors will be taken from the individuals with linkage key information.

For the individuals under subscenarios c and d that are considered linkage error we have to decide to which register the individuals in the CSR would link. We have stated

before that we think that the individuals in the CSR resemble those in the ER most that did not link to the PR. As such, when individuals in the CSR were linkage error, the probability will be higher that they should have been linked to the ER, rather than to the PR. Due to possible errors, some will however have been linked to the PR. It has been chosen that 80 percent will be linked to the ER. The remaining 20 percent will be linked to either the PR or the PR and ER in accordance with the distribution of how the observed data are distributed over the PR or the PR and ER. Given that 60 percent of all observed PR values do not link to the ER and 40 percent do, 60 percent will be linked to the PR and 40 percent will be linked to the PR and the ER.

Note that the individuals will be randomly assigned to link to either the PR, ER or both registers. This will have an effect on the population size estimator. However, this effect will be considerably smaller than randomly assigning all of the 37% of individuals as either linkage error or erroneous captures because the number of observations considered are smaller.

4.5 Analysis

The analysis has been conducted as follows. The dataset used contained all the individuals in the PR, ER and the CSR that did not have a Dutch nationality. We also removed individuals that had a German or Belgium registered place of residence and individuals apprehended by the border police at Schiphol. Usual residence for the PR and ER have been deduced as described in section 3. For the individuals in the CSR that did not link to the PR and ER, usual residence was still missing.

Table 2.1.1. shows that we have 5,087 individuals in the CSR that did not link to the PR and ER. Of these 5,087 individuals, 1,917 individuals (37%) have an incomplete linkage key. Scenario 1 is investigated as follows. From the 1,917 individuals 75% will be considered erroneous captures and 25% will be considered linkage error. First, we remove the 75% of erroneous captures, which entails 1,438 of the 1,917 individuals from the CSR that did not have complete linkage key variables. Then 25% of the individuals in the CSR without complete linkage key variables are considered linkage error, which entailed 479 individuals. As was discussed in section 4.4, we assume 80 percent of the individuals to have linked to the ER, 12 percent to the PR and 8 percent to the intersection of the PR and ER, such that 383 individuals will be linked to the ER, 57 to the PR and 38 to the PR and ER.

Every scenario has a subscenario, where we consider also 5% linkage error and/or erroneous captures from the 3,170 individuals, or 63 percent, that did have complete linkage key values. Assume we took both, and we operate in scenario 1d. Then first, from the 5,087 - 1,917 = 3,170 we delete 5% which are considered erroneous captures, such that 158 individuals are removed from the data. From the 3,170 - 158 = 3,012 individuals remaining, another sample of 5% is considered linkage error. Note that this is 5% of the 3,170 and not 5% of the 3,012 individuals remaining after having already taken 5% erroneous captures. Of these 158 individuals, we link 80 percent, or

126 individuals, to the ER, 12 percent to the PR (19 individuals) and 8 percent to the PR and the ER (13 individuals).

After having dealt with the simulated linkage error and erroneous captures we conduct the multiple imputation using Predictive Mean Matching. We use the package MICE to impute the missing values on the variable usual residence for the CSR using every variable (except the PR) in the imputation model. Mice has been programmed to impute ten times with only one iteration, given that there was only one variable that had missing values.

What follows is carried out for M = 10 imputations and N = 7 nationality groups. First, a loglinear model has been estimated using main effects only. This model is used as a starting point to use the function STEP which selects the best fitting loglinear model on the data. The selection has been done using the BIC. The best fitting loglinear model is used to estimate the size of the population missed. The estimates are stored, averaged over the M = 10 imputations, and reported in this manuscript. The estimate on the total missed portion of the population is found by summing the estimates of all 7 nationality groups. For the two estimates for which a sampling analysis will be done, the analysis also includes a bootstrap with 10,000 iterations to estimate the variance of the resulting estimate. However, we use multiple imputations and for some estimates also multiple samples. We also need to take into account the variance due to both multiple imputation and sampling. The variance for multiple imputations has been coined by Little and Rubin (2012) and an extension from this variance to incorporate variance from multiple samples can be found in the appendix.

5. Results

The results can be found in Table 5.1.1. The maximum likelihood estimates of the missed portion of the population that is estimated by capture-recapture methodology vary considerably over the scenarios. When we consider all 37% of individuals with an incomplete linkage key as belonging to the population that were only registered in the CSR, thus the baseline scenario, the three registers miss a total of 249 thousand individuals. However, when we consider 75% of individuals that have an incomplete linkage key as erroneous captures and the remaining 25% as linkage error the estimate drops to 66 thousand individuals. Thus when we remove 1,438 of the 1,917 individuals as erroneous captures and link the remaining 479 individuals to the other two registers, the resulting population size estimate is rather different from the situation when we consider these 1,917 individuals to belong to the population that are only registered in the CSR. When we consider the full 37% of individuals with an incomplete linkage key as erroneous capture the estimate becomes 151 thousand.

5.1.1 Overview of the scenarios and the resulting maximum likelihood estimates of the missed portion of the population.

	37% with incomplete Linkage key variables		63% with complete linkage key variables		
Scenario	Err. Capt.	Link. error	Err. Capt.	Link. error	Mle
	%	%	%	%	x1000
Baseline	0	0	0	0	249
1a	75	25	0	0	66
1b	75	75	5	0	66
1c	75	25	0	5	54
1d	75	25	5	5	56
2a	100	0	0	0	151
2b	100	0	5	0	151
2c	100	0	0	5	91
2d	100	0	5	5	92

Interestingly, linking an additional 5 percent of the 3,170 individuals with complete linkage key information to the PR and/or ER decreases the population size estimate considerably, whereas removing an additional 5% of erroneous captures from the 3,170 individuals with complete linkage key information does not. From both scenarios it is obvious that linkage error has a bigger effect on the population size estimation than erroneous captures do.

From the scenarios, not considering the baseline scenario, we see that the lowest estimate was 54 thousand and the highest estimate was 151 thousand individuals missed by all three registers. We found that there are 33 thousand usual residents in the ER that were not registered in the PR and thus are part of the undercoverage of the PR. Of the 5 thousand individuals in the CSR that did not link to the PR and ER, approximately 1 thousand individuals are residing longer than a year and have to be considered as part of the under coverage of the PR. Thus, there are 34 thousand usual residents in the undercoverage of the PR that are registered in the ER and CSR and have to be added to the estimates from the analysis to know the undercoverage of the PR.

This means that for the lowest estimate there are 88 thousand individuals in the under coverage of the PR, with a confidence interval of 57 to 151 thousand. For the highest estimate we have 185 thousand individuals not covered by the PR, with a confidence interval of 149 to 222 thousand individuals. Given that there are 16,638 thousand individuals registered in the PR, the undercoverage of the PR of 88 to 185 thousand usual residents means that we have an under coverage of the PR of only .5 to 1.1 percent.

6. Discussion

We are interested in estimating the number of usual residents via capture-recapture analysis to estimate the under count of the PR. We documented in this manuscript the steps taken to overcome the practical challenges that arose during the analyses. As such this manuscript is an example of how the well known capture-recapture methodology can be used in a practical applications such as estimating the undercoverage of a register.

It has to be noted that the data that were used in this manuscript were of individuals aged 15 - 65 years. This was due to restrictions in the registers. The population size estimates and the resulting under coverage of the PR that are described in this manuscript consider only a population of 15 to 65 years of age. To estimate the number of children up until 15 years of age and the elderly over 65 years, other information is needed.

In section 4.2. we found that implied coverages of the PR en the ER are low, which was caused by of the large number of individuals in the CSR that could not be linked. Because of the low implied coverage, the population size estimator will not be robust to possible violations of the assumptions and different scenarios were used to deal with the 37% of individuals in the CSR that had incomplete linkage keys. In the scenarios we randomly assigned cases as either erroneous captures or linkage error because there is no information to know which case belongs to the population and which does not. Unfortunately this random sampling does have an impact on the population size estimate and the variance of the estimate. The confidence interval for our highest estimate becomes as high as 185 thousand. This means that we have an under coverage of the population aged 15 to 65 of maximally 1.1 percent. Even though the confidence interval is high, the under coverage remains rather small.

In Gerritse, Bakker and van der Heijden (2015) we determined a range of possible outcomes, where we expect the population size estimate would lie within. The range of maximum likelihood estimates given in this manuscript includes the lower part of this range of possible outcomes of 175 to 225 thousand individuals. This range of possible outcomes has been created on outcomes of previous research on foreigners and their residence duration. A couple of assumptions had to be made, given that these previous research did not always had the information needed, such as residence duration (Hoogteijling, 2002) or only had information on part of the data (van der Heijden, Cruyff and van Gils, 2011) Thus, even though it is interesting to note that the ranges overlap, implications from this are not straightforward.

It is interesting that the range of possible outcomes of previous research overlaps with the current estimates of 88 to 185 thousand individuals. This may indicate that there is higher possibility that the actual number of usual residents missed by the PR lies close to the upper end of the range of 88 to 185 thousand individuals. If the actual number of under coverage is in the upper end of our range, there is only about

a 1 percent undercoverage of the PR of the population of Dutch usual residents aged 15 to 65.

Extra precautions have been taken to make sure all assumptions have been met. Also, our estimates overlap with a range of possible outcomes by former research. However, it has to be acknowledged that our population size estimates may be biased still, due to unknown violations.

In this manuscript we have given an overview of what was deemed best for the data used at Statistics Netherlands to estimate the under coverage of the PR. The results and their implications do not only apply to the data used in this manuscript, it can be used as a caution for other research as well, especially the more similar their data is to this research. It has to be kept in mind that the capture-recapture methodology can be very sensitive to small changes in the data, but also in the estimation process. It is a useful method yet has strict assumptions which have to be taken into account, but it also becomes more complex when missing data is introduced. In this manuscript we have given our view on a possibility to work with this.

7. Appendix

For the maximum likelihood estimates the analysis also included a bootstrap with 10,000 iterations to estimate the variance of the resulting estimate. This bootstrapped variance is used as the within variance to estimate the total variance for multiple imputation, which is the left hand side of equation (11). In a last loop the stored estimates and variance are used to estimate the variance via (11) This variance is used to derive a 95% confidence interval on the two chosen estimates.

For the estimate of 151 thousand individuals missed (scenario 2a), all individuals with 37% of incomplete linkage key are considered erroneous captures. Thus there is no random assignment of erroneous captures or linkage error and we do not take extra samples to account for variability due to random assignment. For that scenario the variance needed for the confidence interval is estimated via the variance estimation formula of Little and Rubin (2002) that takes into account variance due to multiple imputations.

However, for the estimate of 54 thousand individuals (scenario 1c) there is random assignment of both the linkage error and erroneous captures of the 37% of individuals that have incomplete linkage keys, as well as an additional 5% of linkage error for the individuals in the CSR that did have a complete linkage key. In this manuscript we have multiple sources that affect the variance of the estimate. To assess the variance from multiple imputation, Little and Rubin (2002) have formulated a total variance that combines the within and between imputation variance. Thus the variance for multiple imputation can be accounted for. However, because we also take different samples of erroneous captures and linkage errors we want to account for that variance as well. We formulate here how the total variance formula from Little and Rubin (2002) can be extended to a third source of variance.

Consider the situation where we have taken a sample d of records, to simulate removing erroneous captures. For this sample we perform i=1,…,M imputations. Per imputation we can calculate the capture-recapture estimator $\hat{\theta}_d^i$. An estimate of the variance of $\hat{\theta}_d^i$ conditional on the taken sample, is obtained by applying a parametric bootstrap resampling procedure. We will denote this variance by

$$W_d^i = \widehat{\mathrm{Var}}\left(\hat{\theta}_d^i\right) \tag{6}$$

The final estimator for sample d is averaged over all imputations:

$$\hat{\theta}_d = \frac{1}{M}\sum_{i=1}^{M} \hat{\theta}_d^i \tag{7}$$

Applying the formula for the total variance of $\hat{\theta}_d$ in case of multiple imputation as given e.g. in Little and Rubin (2002) yields

$$\widehat{\mathrm{Var}}\left(\hat{\theta}_d\right) = \frac{1}{M}\sum_{i=1}^{M} W_d^i + \left(1 + \frac{1}{M}\right)\frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_d^i - \hat{\theta}_d\right)^2 \tag{8}$$

We thus have an estimator of the variance for the estimate of θbased on a single sample, with M imputations. Denote this variance by

$$W_d = \widehat{\mathrm{Var}}(\hat{\theta}_d) \tag{9}$$

To be able to estimate the variance due to sampling records to simulate erroneous captures, we will replicate the procedure D times. That is, we will perform the sampling D times and for each sample d=1,…,D we will calculate $\hat{\theta}_d$and its variance estimateW_d. The total variance we will then estimate by applying the general total variance formula

$$\mathrm{Var}\, X = E\big(\mathrm{Var}(X|Y)\big) + \mathrm{Var}\,(E(X|Y)) \tag{10}$$

i.e., the sum of the within-variance and the between-variance. Conditioning on the sampling, we get

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \tfrac{1}{D}\textstyle\sum_{d=i}^{D} W_d + \tfrac{1}{D-1}\sum_{d=1}^{D}\big(\hat{\theta}_d = \bar{\theta}_d\big)^2 \tag{11}$$

Finally, $\bar{\theta}_D = \frac{1}{D}\sum \hat{\theta}_d$can be rewritten into

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{D}\sum_{d=1}^{D}\left(\frac{1}{M}\sum_{i=1}^{M}\widehat{\mathrm{Var}}(\hat{\theta}_d^i)\right) + \frac{1}{D}\sum_{d=1}^{D}\left(\left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_d^i - \hat{\theta}_d\right)^2\right) + \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \bar{\theta}_D\right)^2$$

8. Reference list

Arts, K., Bakker, B. F. M., and Van Lith, E. (2000). Linking administrative registers and household surveys. Special issue Netherlands Official Statistics, 15, 16 - 22.

Bakker, B. F. M. (2009). Trek alle registers open! (Open all registers!) Vrije Universiteit Amsterdam.

Bell, W. R. (1993). Using information from demographic analysis in post-enumeration survey estimation. Journal of the American Statistical Association, 88, 1106 - 1118.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). Discrete multi-variate analysis. MIT press, Cambridge, MA.

Boden, L., I. (2014). Capture-recapture estimates of the undercount of work-place injuries and illnesses: Sensitivity analysis. American Journal Of Indus-trial Medicine, 57, 1090 - 1099.

Brown, J. J., Abbott, O., and Diamond, I. D. (2006). Dependence in the 2001 one-number census project. Journal of the Royal Statistical Society. Series A (Statistics in Society), 169, 883 - 902.

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999). A methodological strategy for a one-number Census in the UK. Journal of the Royal Statistical Society. Series A, 162:247 - 267.

Buuren, van, S. (2012). Flexible imputation of missing data. Chapman & Hall/CRC, Boca Raton, FL.

Cormack, R. M. (1989). Log-linear models for capture-recapture. Biometrics, 45, 395 - 413.

European Parliament (2008). Regulation (ec) no 763/2008 of the european parliament and the council of 9 july 2008 on population and housing censuses. Official Journal of the European Union, 13.8.2008:L 218/14 - L 218/20.

Fellegi, I. P. and Sunter, A. B. (1969). A theory of record linkage. Journal of the American Statistical Association, 64, 1183 - 1210.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. Biometrika, 59, 409 - 439.

Gerritse, S. C., Bakker, B. F. M., and van der Heijden, P. G. M. (2015). Different methods to complete datasets used for capture-recapture estimation: estimating the number of usual residents in the Netherlands, Statistical Journal of the IAOS, 31, 613–627.

Gerritse, S. C., Bakker, B. F. M., Zult, D., and van der Heijden, P. G. M. (Submitted). The effects of imperfect linkage and erroneous captures on the population size

estimator. Survey Methodology.

Gerritse, S. C., Van der Heijden, P. G. M., and Bakker, B. F. M. (2015). Sensitivity of population size estimation for violating parametric assumptions in loglinear models. Journal of Official Statistics, 31, 357 - 379.

Herzog, T. N., Scheuren, F. J., and Winkler, W. (2007). Data Quality and Record Linkage Techniques. Springer-Verlag, New York.

Hoogteijling, E. J. M. (2002). Raming van het aantal niet in de gba geregistreer den. Technical report, Centraal Bureau voor de Statistiek, Voorburg / Heerlen.

International Working Group for Disease Monitoring and Forecasting (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. American Journal of Epidemiology, 142, 1047 - 1058.

Lanzieri, G. (2013). Population definitions at the 2010 censuses round in the countries of the UNECE region. 5th meeting of the UNECE Group of Experts on Population and Housing Censuses, Geneva, 30 September - 3 October 2013.

Little, R. J. and Rubin, D. B. (2002). Statistical analysis with missing data. John Wiley and Sons, Hoboken, New Jersey.

Nirel, R. and Glickman, H. (2009). Sample surveys and censuses. Sample surveys: Design Methods and Applications, 29A, 539 - 565.

ONS (2012). The 2011 Census coverage assessment and adjustment process. Office for National Statistics, Newport, South Whales.

Sekar, C. C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. Journal of the American Statistical Association, 44, 101- 115.

United Nations Statistics Division (2008). Principles and Recommendations for Population and Housing Censuses - Rev.2. Statistical papers Series M No 67/Rev.2. United Nations, New York.

Van der Heijden, P. G. M., Whittaker, J., Cruyff, M. J. L. F., Bakker, B. F. M., and Van der Vliet, H. N. (2012). People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. The Annals of Applied Statistics., 6,831 - 852.

Wolter, K. M. (1986). Some coverage error models for census data. Journal of the American Statistical Association, 81,338 - 346.

Zhang, L.-C. (2015). On modelling register coverage errors. Journal of Official Statistics, 31:381-396.

Zhang, L.-C. and Dunne, J. (2015). Population size estimation based on administrative registers. 4th Baltic-Nordic Conference on Survey Statistics Conference Proceedings, 24 - 28 August 2015, Helsinki.

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/'15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/'13–2014/'15	Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colofon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Studio BCO

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information